2018-04-29

Classification of Bone Cements Using Multinomial Logistic Regression Method

Jinglun Wei
Worcester Polytechnic Institute

Classification of Bone Cements Using Multinomial Logistic Regression Method

by

Jinglun Wei

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

by

April 2018

APPROVED:

Professor Thelge Buddika Peiris, Major Thesis Advisor

Abstract

Bone cement surgery is a new technic widely used in medical field nowadays. In this thesis I analyze 48 bone cement types using their content of 20 elements. My goal is to find a method to classify new found bone cement sample into these 48 categories. Here I will use multinomial logistic regression method to see whether it works or not. Due to the lack of observations, I generate enough data by adding white noise in proper scales to the original data again and again, and then I get a data set of over 100 times as many points as the original one. Then I use purposeful variable selection method to pick the covariates I need, rather than stepwise selection. There are 15 covariates left after the selection, and then I use my new data set to fit such a multinomial logistic regression model. The model doesn't perform that good in goodness of fit test, but the result is still acceptable, and the diagnostic statistics also indicate a good performance. Combined with clinical experience and prior conditions, this model is helpful in this classification case.

Acknowledgements

I would like to express my gratitude to my advisor, who encouraged me many times and gave me lots of suggestions on this hard work.

Thanks also to my wife, the fact she was by my side all the time while I was writing this thesis.

Contents

List of Tables

Chapter 1

Introduction.

Bone cement is a class of synthetic organic or inorganic materials and is now widely used in the medical field. It does not glue adjacent bones but acts as a "grout" to create a tight space and fill up the cavity that holds the bones together (see Vaishya et al., 2013). However, except for the medical field, bone cement may have more advanced use in some other fields. For example, if a corpse is found hard to identify the identity but he or she has had a surgery implanting bone cements, one can measure the content of the elements in his or her bone cements. It sounds good if one can tell something from the bone cements of the corpse, and then one may have the information like the dead person's name, which hospital he or she did this surgery and so on. In my project, I've got 48 bone cement types and marked by different IDs, and I measured the content of 20 elements which are K, Ti, V, Cr, Mn, Ni, Cu, Zn, As, Se, Br, Rb, Sr, Hg, Ga, Pb, Nd, Lu, Ta and Ln in each cement type. Most of the cement types have three observations, and a few have two or just one observation. So totally, I have 141 observations. By using all the information, I will aim to give the criteria to classify a new given sample into one of the 48 types.

The model I decide to use is multinomial logistic regression. Logistic regression

model is a very useful model when describing the relationship between discrete outcome variables, talking on at least two possible values, and one or more predictor variables. The goal of this model, which is the same as that of any other regression method, is to find the best fitting and most convincing and clinically interpretable model to describe such a relationship (see David et al., 2013). What makes a difference between a logistic regression model and a linear regression model is the form of outcome variables, reflecting in the form of model and the assumptions. However, once this difference is taken into account, linear regression and logistic regression share many similar ideas, such as the general principles and the analysis technics.

This thesis is organized as follows: A brief overview of binary and multinomial logistic regression is given in the first part of chapter 2. Then later in this chapter, I introduce a test to check the goodness of overall fit and also four diagnostic statistics to find out poorly performed and influential individuals. In chapter 3, I build a multinomial logistic regression model by fitting the bone cement data, and assess the model using the technics I just mentioned. Then in the last chapter, I give a discussion on the final model including some good results and an expectation for future work.

Chapter 2

Multinomial logistic regression.

2.1 Model.

2.1.1 Binary logistic regression.

As mentioned before, a binary logistic regression model can explain the relationship between a dichotomous outcome variable and one or more covariates. Suppose that the outcome variable is denoted by Y and the covariates are denoted by a vector $\mathbf{X}$ with elements X_js, where Y has two possible values 0 and 1 and j=1, 2, $\cdots$, p. The key quantity in any regression problem is the mean value of the outcome variable, given the values of the covariates. This quantity can be expressed as $E(Y|\mathbf{x})$, where $\mathbf{x}$ denotes $\mathbf{X}$ taking all elements specific values. To simplify the work, I use $\pi(\mathbf{x})$ to express this conditional mean. Thus the form of binary logistic regression model is:

$$\pi(\mathbf{x}) = \frac{\mathrm{e}^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+\mathrm{e}^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}, \tag{2.1}$$

where β_0 is the intercept and β_1, β_2, $\cdots$, β_p are the coefficients of x_1, x_2, $\cdots$, x_p, respectively. For better use of the properties of linear regression model, a transfor-

mation of $\pi(\mathbf{x})$ has been defined as:

$$\begin{aligned} g(\mathbf{x}) &= \ln[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}], \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \end{aligned} \tag{2.2}$$

$g(\mathbf{x})$ is called the logit and is linear in parameters. It may be continuous and range from $-\infty$ to ∞ depending on the value of $\mathbf{x}$.

Moreover, people assume that an observation y is expressed as $y = \pi(\mathbf{x}) + \epsilon$, where ϵ denotes the deviance between the observation and the condition mean. It is clear that ϵ may take one of two possible values, when $\mathbf{x}$ is given. If $y = 0$, $\epsilon = -\pi(\mathbf{x})$ with probability $1 - \pi(\mathbf{x})$, and if $y = 1$, $\epsilon = 1 - \pi(\mathbf{x})$ with probability $\pi(\mathbf{x})$. Therefore, ϵ has a distribution with mean 0 and variance $\pi(\mathbf{x})[1 - \pi(\mathbf{x})]$. That is the outcome variable Y follows a binomial distribution with probability $\pi(\mathbf{x})$, when given $\mathbf{x}$.

2.1.2 Multinomial logistic regression.

Multinomial logistic regression model is an extension of the binary one, which means the outcomes have multiple levels but not dichotomous. It is noteworthy that when talking about a discrete-outcome regression model with at least three responses, the measurement scale should be taken into consideration. In my case, the outcomes are in nominal scale, where there is also ordinal scale correspondingly. Now suppose that my outcome variable Y has a categories, which are coded from 1 to a, and I have p predictor variables. Unlike in binary case, a multinomial model requires to set a baseline at first. For instance, I can use $Y = 1$ as the baseline and then form $a-1$ logit functions, where the natural logarithm of the odds is expressed

by a linear combination of all covariates and a constant term, which are denoted as:

$$\begin{aligned} g_i(\mathbf{x}) &= \ln[\frac{\Pr(Y=i|\mathbf{x})}{\Pr(Y=1|\mathbf{x})}], \\ &= \beta_{i0} + \beta_{i1}x_1 + \cdots + \beta_{ip}x_p, \end{aligned} \tag{2.3}$$

where $\mathbf{x}$ is a vector of p covariates with elements x_1, x_2, $\cdots$, x_p, β_{i0} is the intercept, β_{i1}, β_{i2}, $\cdots$, β_{ip} are the coefficients of the covariates, respectively, and i=2, 3, $\cdots$, a. So the conditional probabilities of each category given the observed values of the covariates are

$$\begin{aligned} \pi_i(\mathbf{x}) &= \Pr(Y=i|\mathbf{x}), \\ &= \frac{\mathrm{e}^{g_i(\mathbf{x})}}{\sum_{k=1}^{a}\mathrm{e}^{g_k(\mathbf{x})}}, \end{aligned} \tag{2.4}$$

where i=1, 2, 3, $\cdots$, a, and $g_1(\mathbf{x}) = 0$. Then I utilize the maximum likelihood method to estimate the parameters. To construct the likelihood function, I create a binary response variables Y_1 to Y_a, which are coded as follows: if $Y = i$, then $Y_i = 1$ and $Y_{s\neq i} = 0$, where i, s=1, 2, $\cdots$, a. So no matter what value Y takes, the sum of these a variables is always $\sum_{i=1}^{a} Y_i = 1$. Assume that there are n independent observations, so the likelihood function is

$$L(\beta) = \prod_{k=1}^{n}[\pi_1(\mathbf{x}_k)^{y_{1k}}\pi_2(\mathbf{x}_k)^{y_{2k}}\cdots\pi_a(\mathbf{x}_k)^{y_{ak}}], \tag{2.5}$$

where k=1, 2, $\cdots$, n. Thus taking the log and using the fact that $\sum_{i=1}^{a} Y_{ik} = 1$ for each k, the log likelihood function is

$$l(\beta) = \sum_{k=1}^{n}[y_{2k}g_2(\mathbf{x}_k) + \cdots + y_{ak}g_a(\mathbf{x}_k) - \ln(1 + \mathrm{e}^{g_2(\mathbf{x}_k)} + \cdots + \mathrm{e}^{g_a(\mathbf{x}_k)})]. \tag{2.6}$$

Then by taking the first partial derivatives of $l(\beta)$ with respect to each β_{ij}, where i=1, 2, $\cdots$, n and j=1, 2, $\cdots$, p, I can obtain the maximum likelihood estimators.

Before taking inferences from the fitted model, the overall fit and the contribution of each individual observation to the fit should be assessed, where the multiple outcome levels make it more complex than in binary cases.

2.2 Goodness of fit.

To begin with, in binary cases, the decile of risk goodness of fit test can be used to check the performance of overall fit, which divides all the observations into several groups and then constructs a statistic following a chi-square distribution. For multinomial logistic regression, Fagerland (2009) and Fagerland et al. (2008) developed an extension of this test. The extension test forms g groups using the ranked values of $1-\hat{\pi_1}$ that is the complement of the fitted value of $\Pr(Y=1|\mathbf{x})$, and then a table of observed and estimated expected frequencies within the a outcome levels and g groups to assess departures from model fit. So the test statistic is

$$\hat{C_M} = \sum_{s=1}^{g}\sum_{i=1}^{a}\frac{(O_{si}-\hat{E}_{si})^2}{\hat{E}_{si}}, \tag{2.7}$$

where $O_{si} = \sum_{l\in\Omega_s} y_{li}$, $\hat{E}_{si} = \sum_{l\in\Omega_s}\hat{\pi_{li}}$, and Ω_s denotes the observations in the sth group. If the sample size is large enough and the fitted model is proper, the statistic shown above should follow a chi-square distribution with $(g-2)\times(a-1)$ degree of freedom.

2.3 Diagnostics.

Next, it is important to check for individuals performing poor and influential. Although my model is multinomial, I should check the diagnostic statistics by each single logit function, so the multinomial case becomes the same as the binary case. For linear regressions, the key quantities for diagnostics are the residual sum-of-squares, but there is a little difference between the linear and logistic case as I indicated before, where the errors are binomial in logistic regressions. Thus I introduce two kind of residuals, which is not simply the difference between fitted values and observed values. The first one is Pearson residual, which is expressed as:

$$r(y_i, \hat{\pi}_i) = \frac{(y_i - \hat{\pi}_i)}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}, \tag{2.8}$$

and the other one, Deviance residual, has an expression as:

$$d(y_i, \hat{\pi}_i) = \pm[2[y_i \ln(\frac{y_i}{\hat{\pi}_i}) + (1 - y_i) \ln(\frac{1 - y_i}{1 - \hat{\pi}_i})]]^{\frac{1}{2}}. \tag{2.9}$$

Then based on the knowledge of these two residuals, I introduce four helpful diagnostic statistics, which are h_i, $\Delta\hat{\beta}_i$, ΔX_i^2 and ΔD_i.

Firstly, a linear approximation to the fitted values can be derived by using the weighted least squares linear regression as a model, and will yield a hat matrix $\mathbf{H}$ for logistic regression. The matrix is expressed as:

$$\mathbf{H} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{\frac{1}{2}}, \tag{2.10}$$

where $v_i = \hat{\pi}(\mathbf{x}_i)[1 - \hat{\pi}(\mathbf{x}_i)]$ is the element of the diagonal matrix $\mathbf{V}$. I use h_i to denote the ith diagonal element of the matrix $\mathbf{H}$, which is proportional to the

distance between $\mathbf{x}$ and the mean $\bar{\mathbf{x}}$ and is named leverage value.

The other three statistics are from the idea that removing one observation one time and then checking the effect on estimated coefficients and overall measures of fit $\mathbf{X}^2$ and $\mathbf{D}$, and are expressed as:

$$\Delta\hat{\beta}_i = \frac{r_i^2 h_i}{(1-h_i)^2}, \tag{2.11}$$

$$\Delta X_i^2 = \frac{r_i^2}{1-h_i}, \tag{2.12}$$

$$\Delta D_i = \frac{d_i^2}{1-h_i}. \tag{2.13}$$

Larger values of ΔX_i^2 or ΔD_i indicate the poorly fitted observations and a higher value of $\Delta\hat{\beta}_i$ shows that the observation has great influence on the estimated parameters.

Chapter 3

Analysis.

As introduced above, the original data set I use has 141 observations, 20 covariates and 48 outcome categories. Due to the cost and difficulty, these 141 observations are all I have, although there seems to be too few observations to keep the model stable. Therefore, if more information could be reached in future work, the model and inferences might be better and more accurate. To build the multinomial logistic regression model, I start with all the 20 covariates together and do purposeful selection of these covariates first. Hence, I begin by fitting the 20 univariable models separately, and the results are shown in table 3.1.

According to the likelihood ratio test of every univariable model, the p-values of the variables Se, Pb, Lu, Ta and Ln are not significant at the 0.25 level, so I eliminate these five variables from the model. Then I use the other 15 variables to construct my first multinomial logistic model, and also use the likelihood ratio test to check if any independent variable is not significant. The results are not that good and some of the coefficients are insignificant, since the observations are too few leading to large standard errors. Thus I decide to generate more observations by adding white noise in proper scales to the original data time by time and finally

Variable	Chi-square	DF	p-value
K	367.45	47	<0.001
Ti	307.15	47	<0.001
V	269.66	47	<0.001
Cr	365.94	47	<0.001
Mn	405.27	47	<0.001
Ni	293.4	47	<0.001
Cu	255.04	47	<0.001
Zn	278.97	47	<0.001
As	126.43	47	<0.001
Se	22.349	47	0.9991
Br	231.46	47	<0.001
Rb	190.51	47	<0.001
Sr	356.37	47	<0.001
Hg	73.702	47	0.007686
Ga	100.25	47	<0.001
Pb	46.016	47	0.5133
Nd	100.65	47	<0.001
Lu	21.188	47	0.9996
Ta	8.0713	47	1
Ln	24.285	47	0.9975

Table 3.1: Results of Fitting Univariate Models with the 20 IDs as the Outcome

get 15228 observations, showing that the data set is enlarged by 108 times. Now I refit my model with these points and calculate the p-values from the Wald tests. The results show the variables are significant in most of the logit functions, where there are 47 logit functions by setting one outcome category as the baseline, but are not significant in the other functions. Here I pick out the variables As, Br, Rb, Hg, Ga and Nd, which perform worse than the other variables, and some results are shown in table 3.2.

However, this time I do not eliminate all these variables directly, because they may give information to some extent. These six variables are checked one by one, by comparing the model with the variable and the one without the variable, and if the coefficients for the remaining variables are changed by more than 20-25%, the

logits	As	Br	Rb	Hg	Ga	Nd
104	0.403	0.183	0.039	<0.001	0.793	0.807
13-29	0.428	0.009	0.754	0.719	0.457	0.983
133	0.022	0.574	0.958	<0.001	0.005	0.334
14-29	0.209	0.381	0.445	0.002	0.901	0.250
145	0.615	0.960	0.962	0.593	0.993	0.088
15-29	0.528	0.252	0.953	0.782	0.793	0.001
152	0.264	0.131	0.901	0.994	0.778	0.015
16-25	0.898	0.513	0.985	0.858	0.983	0.344
173	0.002	0.843	<0.001	0.372	0.864	0.328
174	0.895	0.706	0.155	0.478	0.971	<0.001

Table 3.2: Part of the Two-Tailed p-Values for the Bad Performed Six Covariates

variable should be kept in our model. Here I show the comparison of the first three logit functions and the first five elements as an example, and results of the other functions are similar. The results are shown in table 3.3.

Although the coefficients for the same covariates between the full model and any reduced model seem not to change much, there is still at least one pair of coefficients for each covariate different by more than 25%. Therefore I decide to hold all these six variables back in my model for the sake of enough information, and now I consider the model containing K, Ti, V, Cr, Mn, Ni, Cu, Zn, As, Br, Rb, Sr, Hg, Ga and Nd as my preliminary main effects model.

Next, before I reach the final model, the linearity of continuous variables and the existence of interactions should be checked. Here I have checked the scale of all the covariates included in the preliminary main effects model by using fractional polynomials. The analysis shows that there is no evidence of nonlinearity in every logit function for these variables, so I do not need to do any transformation on our variables, such as taking the square root or the log. As for the interactions, I assume that there are no interaction items for two reasons: one is most of the elements I discussed at the beginning are rare metals and I do not use the compounds of the

Logits	K	Ti	V	Cr	Mn
104	-50.9069470	-365.16169	10.91801	-531.42826	-298.769502
13-29	18.8517425	-532.41428	-304.77829	-27.15587	1440.631910
133	127.8939807	674.29214	-63.04616	-230.61843	799.575602
104	-43.9233629	-352.699579	11.78747	-550.29599	-319.092256
13-29	29.7672671	-563.615243	-317.47313	-29.77381	1512.177112
133	112.9848219	718.062966	-63.83036	-201.12271	830.973364
104	-38.936587	-397.0772720	-36.471537	-546.25814	-313.507926
13-29	10.603553	-536.2941278	-307.561501	-43.63762	1462.656741
133	161.593428	691.1899550	-57.622457	-164.69359	784.156116
104	-34.247842	-395.01393	13.04883	-507.56924	-258.279686
13-29	26.164373	-530.69359	-321.47240	-41.89679	1451.389846
133	111.986593	682.14886	-60.69639	-178.37324	771.945079
104	-54.470580	-445.35275	59.47506	-517.10971	-294.818922
13-29	22.780835	-550.42974	-284.78721	-43.56959	1480.439211
133	77.870368	674.56050	-60.35487	-201.85216	813.886951
104	-42.1191835	-384.31817	2.808182	-529.47595	-297.685673
13-29	13.3184047	-532.55623	-302.463563	-30.40720	1422.575845
133	129.7065990	684.25617	-63.200453	-209.84757	784.562272
104	-50.7920686	-352.599193	98.915332	-520.88127	-312.4631939
13-29	8.3730056	-434.337189	-202.198930	-18.68266	1377.3857285
133	129.5928011	649.883050	-73.301099	-173.37182	765.7656425

Table 3.3: Comparison of the Coefficients for the First Three Logit Functions

elements, and the other is the cost of checking the interactions is too high, so I may leave this for further study.

Thus I have got my preliminary final model as in table 3.4, 3.5, 3.6 and 3.7. Table 3.4 and 3.6 combined show the first 22 logit functions, and the estimated values of intercepts and coefficients are listed under *Intercept* and the 15 specific elements, correspondingly. Table 3.5 and 3.7 are the same but show the other 25 logit functions.

Before making inferences, the last step is to assess the goodness of the overall fit and the contribution of each single subject. As introduced before, an extension of the decile of risk test can be used to test the overall goodness of fit for a multinomial logistic regression model. Since my model has 15 covariates, I will set the number

Logits	Intercept	K	Ti	V	Cr	Mn	Ni	Cu
104	41.9	-50.9	-365.2	10.9	-531.4	-298.8	231.8	323.2
13-29	-60.2	18.9	-532.4	-304.8	-27.2	1440.6	-220.9	-209.3
133	-85.3	127.8	674.2	-63.0	-230.6	799.5	-23.5	619.2
14-29	-156.2	307.5	327.5	370.5	-394.6	122.8	193.4	-229.1
145	-33.4	-30.2	527.1	-329.2	-139.2	-903.8	133.027994	-252.5
15-29	-53.9	7.0	531.4	-347.8	-804.0	202.9	664.4	-150.0
152	-15.3	29.3	-24.4	-27.7	-438.6	819.4	-844.3	529.6
16-25	3.9	9.7	-61.7	-205.2	-850.6	-451.7	-378.5	-328.6
173	11.3	-14.3	197.6	458.2	-305.0	-1303.7	184.8	437.2
174	39.0	-55.3	-527.5	-314.4	-978.5	-127.7	-223.8	-75.5
18-29	-145.6	15.9	-68.8	223.3	1806.1	213.0	-284.5	-20.6
19-Apr	-50.6	31.3	157.5	-20.8	-74.1	110.9	-137.0	109.4
19-Jun	22.3	36.6	-256.6	192.3	-445.0	12.1	-305.0	32.1
19-May	-185.6	144.9	1151.0	31.5	-129.9	638.5	-49.7	-162.9
19-Oct	-2.3	21.5	-237.1	-37.7	-76.3	-375.8	15.2	-94.3
192	-167.7	-27.9	1305.8	368.9	-189.2	1132.5	-466.5	176.0
21-25	-127.0	26.8	80.7	289.8	1220.1	88.6	791.1	-103.5
25-32	-141.6	-23.9	-135.5	-17.3	1279.0	58.3	847.9	-18.0
26-29	-83.0	-45.3	-466.5	103.8	1046.9	-71.0	937.5	280.6
27-25	2.3	-3.6	-839.4	69.5	497.4	-203.8	977.9	-173.2
28-25	16.4	12.2	-654.9	-32.6	364.6	-427.7	751.0	-18.3
29-25	-12.6	51.3	-479.3	30.5	228.4	-109.2	569.8	-88.2

Table 3.4: Preliminary Final Model: Part 1

of groups as 30, which is required to be larger than the number of covariates. The results are not that good but acceptable more or less. The chi-square statistic is not small enough resulting in the p-value around 0.25, although it can't reject the null hypothesis that the actual and predicted values are similar across 30 deciles. The reason for this could be lack of observations or overfitting caused by my huge generated data set. As I discussed before, it is not possible to have new data included, so what I can do is very limited. However, it is still very plausible that I may get useful information and inspiration from my model to some extent.

Nevertheless, it is comforting that the model performs well in diagnostics. Since my new data set is too large and derived from the original data set with white

Logits	Intercept	K	Ti	V	Cr	Mn	Ni	Cu
29-Dec	-15.9	13.4	-310.4	-328.5	-163.3	-5.7	-126.4	-71.6
29-Jul	-1.1	30.5	-16.4	482.6	-220.6	46.9	-3.9	-343.4
29-Mar	23.1	-5.0	-98.2	22.1	-227.2	77.7	-182.0	-169.6
30-25	-64.5	32.1	-530.2	-217.3	603.4	1138.1	559.2	-181.8
31-51	-57.6	36.3	-279.4	-817.7	437.9	-597.8	614.9	-12.7
74	-90.9	-3.5	422.0	-649.3	-344.2	-301.4	-339.0	-150.5
93B	-123.6	111.8	71.3	823.1	102.5	8.8	222.8	-82.7
93C	-42.9	121.9	-360.4	228.3	85.6	211.0	-254.4	166.2
93D	-133.5	0.2	96.0	-113.8	-227.3	182.9	-159.0	36.3
94A	-246.3	487.4	372.7	69.7	39.2	-153.4	-37.0	70.0
94C	-61.9	-0.2	-145.6	-49.5	376.8	-182.1	321.4	14.2
Aug-32	-2.0	68.4	-679.6	-499.1	-292.7	-655.9	-13.9	-155.7
B	-24.7	21.6	106.2	138.0	64.7	219.1	-394.5	-103.0
CA	18.3	-75.5	-198.7	180.5	890.5	-417.4	-106.6	-52.0
CB	18.1	-16.8	81.8	193.3	604.6	-840.1	-493.0	-267.6
CC1	-14.5	-35.4	302.5	457.2	628.0	26.5	-292.0	147.7
CC2	-11.3	58.6	248.0	352.4	432.8	-73.3	135.8	-162.5
CHILEAN	-61.5	-3.5	403.4	-26.6	-264.9	584.2	-302.9	347.3
G	42.45	14.0	-620.0	-126.5	49.3	492.3	145.3	366.0
Jan-95	22.0	76.2	-483.1	-187.3	-798.6	-266.6	329.6	320.2
L	43.3	-130.9	178.1	-1084.1	-358.5	-1311.1	-366.3	126.3
Portland	-71.8	6.5	143.9	203.9	200.0	1083.1	-612.7	-181.1
T1-LE	-46.0	-58.8	-61.6	47.7	91.6	44.9	73.4	10.9
T1-LF	-20.3	9.9	456.4	35.5	-93.6	-62.9	-487.5	302.1
T3	-112.9	38.4	819.4	230.5	-245.5	848.8	-63.2	-96.3

Table 3.5: Preliminary Final Model: Part 2

noise in different proper scales, it is feasible to use the original data set to check all observations, so that time and cost are saved. No large values of the four statistics, h_i, $\Delta\hat{\beta}_i$, ΔX_i^2 and ΔD_i, have been found. Therefore I can do some inferences carefully with clinical experience then.

Logits	Zn	As	Br	Rb	Sr	Hg	Ga	Nd
104	305.2	255.3	-188.2	429.4	-908.5	435.7	-20.8	48.5
13-29	112.4	-30.0	-181.7	-7.4	587.4	-38.4	19.5	6.5
133	-108.7	80.6	29.9	1.3	-1042.1	-290.9	25.3	363.3
14-29	-101.8	-57.3	298.5	-102.5	586.2	-103.7	11.5	308.7
145	298.7	-67.5	-13.4	6.8	309.9	-69.7	2.6	-174.3
15-29	424.9	-192.7	254.3	8.7	-10.9	-100.3	84.7	-323.5
152	-384.1	164.2	146.8	-11.6	-701.9	-1.1	-15.6	370.5
16-25	-475.0	-48.3	157.4	8.1	420.2	-52.5	7.4	59.2
173	201.1	105.5	-10.9	-107.8	-631.6	39.0	4.9	-206.7
174	758.6	-45.7	64.0	-98.8	-790.1	-91.0	-3.3	-585.2
18-29	54.3	37.9	39.9	81.9	105.5	12.4	28.0	-173.9
19-Apr	255.6	235.2	-364.2	270.6	653.0	-33.6	-101.8	120.5
19-Jun	-854.5	-204.4	-229.9	-38.1	-0.8	-4.1	-138.0	-53.2
19-May	-294.6	18.4	194.3	14.3	451.9	-184.4	-17.5	235.0
19-Oct	-65.8	-71.0	-314.4	107.5	581.3	86.6	104.3	-107.6
192	79.4	54.0	-26.1	43.7	9.1	11.8	103.4	-65.5
21-25	129.9	45.8	88.8	-101.3	-199.5	37.3	46.6	-10.8
25-32	-409.2	-38.2	14.8	65.8	640.6	4.0	-0.3	62.7
26-29	-276.2	-108.3	-93.9	-85.2	-508.6	-55.1	-85.3	375.0
27-25	153.7	46.9	2.2	-4.0	-242.5	-20.8	-2.9	-43.5
28-25	342.6	-63.5	49.3	104.6	-221.3	100.2	-6.3	-123.7
29-25	-41.5	70.7	302.3	70.2	396.9	-14.7	-8.7	0.4

Table 3.6: Preliminary Final Model: Part 3

Logits	Zn	As	Br	Rb	Sr	Hg	Ga	Nd
29-Dec	29.5	-11.9	499.2	270.7	700.9	-20.2	99.5	-146.2
29-Jul	1107.9	11.7	62.6	-35.5	-1242.1	51.7	-40.9	107.3
29-Mar	257.7	-75.7	-154.0	7.0	-767.6	-166.4	18.3	117.8
30-25	-53.5	-10.7	26.3	-14.3	452.3	30.1	-24.2	162.2
31-51	292.8	-38.3	-128.1	-2.7	915.4	16.8	-4.1	-163.0
74	981.1	-88.2	-205.9	27.3	933.5	60.6	5.3	-60.6
93B	23.4	-0.4	-112.2	-5.5	-335.9	47.1	-49.2	173.6
93C	-142.6	-0.3	263.3	1.0	325.2	-49.5	98.5	-175.2
93D	-275.9	-36.7	250.8	6.9	963.0	9.0	-66.4	-448.6
94A	-167.1	-1.3	-27.7	26.2	759.9	247.5	-8.9	145.0
94C	262.7	64.1	-230.0	-11.8	800.3	-2.6	49.0	14.4
Aug-32	-966.2	17.6	206.9	-73.9	903.6	30.2	109.8	-399.5
B	-233.0	-195.6	-258.2	-288.1	321.3	1.2	75.7	-287.1
CA	-583.8	-11.7	-220.6	47.9	-605.4	72.2	63.7	224.5
CB	-435.3	-132.5	107.1	-82.6	-705.0	44.7	123.2	-552.3
CC1	-596.3	159.8	78.3	-38.9	-487.1	-115.4	-121.7	6.8
CC2	-647.9	41.4	-211.3	3.9	-772.4	-14.4	-48.2	324.6
CHILEAN	-1025.4	111.8	-232.6	-51.2	562.1	11.6	-9.8	-2.7
G	53.1	115.8	-279.3	-47.8	-1161.1	21.8	11.8	-165.7
Jan-95	-350.0	13.3	38.1	30.0	-6.4	-78.9	-67.3	-307.0
L	-1766.9	-162.0	-159.1	-245.6	-240.2	-79.9	-124.1	615.3
Portland	769.2	-84.0	-173.8	27.4	16.1	-5.5	-70.1	41.5
T1-LE	-74.4	5.4	60.6	1.2	200.8	6.9	1.5	417.4
T1-LF	-19.5	-8.8	45.7	-41.3	-424.6	84.6	-87.4	372.3
T3	925.4	8.0	-89.4	-66.9	-102.4	36.7	-48.0	-284.7

Table 3.7: Preliminary Final Model: Part 4

Chapter 4

Conclusion and discussion.

Now I have reached my final model, which was the so-called preliminary final model. I can see how my model works or performs by refitting the original data set back, and check the probability in each category. All the results are listed in table 4.1, 4.2, 4.3 and 4.4. The following tables show the prediction probabilities of each observation in its actual category and predicted category. If the predicted type for an observation is the same as the actual type, meaning that the observation is predicted correctly, there will be a dash in the last column.

Since there are 141 observations totally in the original data set, the correct rate is around 90%. Also, the wrong arrangement occurs only in several specific categories, so when I work practically, I can have enough confidence to say the bone cement is in one of the other categories or in these categories with some clinical experience or prior conditions.

There are still problems can be addressed in future study, such as collecting more observations and finding out why the model performs not that good in the goodness of fit test. I can try some other classification methods as well like random forest, support vector machines and so on. Here I just give some inspiration through the

observation	actual type	prob	predicted type	prob
1	Portland	0.999	Portland	-
2	Portland	0.999	Portland	-
3	Portland	0.986	Portland	-
4	CA	0.039	CC1	0.482
5	CA	0.403	CA	-
6	CA	0.264	CC2	0.420
7	CB	0.061	CC2	0.911
8	CB	0.287	CC2	0.304
9	CB	0.126	CC2	0.760
10	CC1	0.628	CC1	-
11	CC1	0.386	CC2	0.556
12	CC1	0.372	CA	0.510
13	CC2	0.434	CC2	-
14	CC2	0.567	CC2	-
15	CC2	0.635	CC2	-
16	L	0.531	L	-
17	L	0.005	19-Jun	0.614
18	L	0.645	L	-
19	G	0.932	G	-
20	G	0.874	G	-
21	G	0.940	G	-
22	B	0.989	B	-
23	B	0.907	B	-
24	B	0.969	B	-
25	T3	0.958	T3	-
26	T3	0.899	T3	-
27	T3	0.885	T3	-
28	T1-LF	0.897	T1-LF	-
29	T1-LF	0.818	T1-LF	-
30	T1-LF	0.839	T1-LF	-
31	T1-LE	0.912	T1-LE	-
32	T1-LE	0.991	T1-LE	-
33	T1-LE	0.840	T1-LE	-
34	74	0.966	74	-
35	74	0.881	74	-
36	74	0.942	74	-
37	93B	0.831	93B	-

Table 4.1: Original Observations Rearrangement: Part 1

observation	actual type	prob	predicted type	prob
38	93B	0.957	93B	-
39	93B	0.827	93B	-
40	94A	0.869	94A	-
41	94A	0.854	94A	-
42	94A	0.892	94A	-
43	173	0.822	173	-
44	173	0.987	173	-
45	173	0.986	173	-
46	174	0.931	174	-
47	174	0.85	174	-
48	174	0.055	29-Mar	0.788
49	CHILEAN	0.999	CHILEAN	-
50	CHILEAN	0.999	CHILEAN	-
51	CHILEAN	0.999	CHILEAN	-
52	192	1.000	192	-
53	192	1.000	192	-
54	192	1.000	192	-
55	94C	1.000	94C	-
56	94C	1.000	94C	-
57	94C	1.000	94C	-
58	104	0.964	104	-
59	104	0.923	104	-
60	104	0.909	104	-
61	93D	1.000	93D	-
62	93D	1.000	93D	-
63	93D	1.000	93D	-
64	93C	0.999	93C	-
65	93C	0.997	93C	-
66	93C	0.999	93C	-
67	145	0.732	145	-
68	145	0.430	145	-
69	145	0.877	145	-
70	152	0.675	152	-
71	152	0.654	152	-
72	152	0.860	152	-
73	133	0.906	133	-
74	133	0.598	133	-

Table 4.2: Original Observations Rearrangement: Part 2

observation	actual type	prob	predicted type	prob
75	133	0.886	133	-
76	Jan-95	0.953	Jan-95	-
77	Jan-95	0.502	Jan-95	-
78	Jan-95	0.730	Jan-95	-
79	1-Feb	0.939	1-Feb	-
80	1-Feb	0.796	1-Feb	-
81	1-Feb	0.842	1-Feb	-
82	29-Mar	0.569	29-Mar	-
83	29-Mar	0.592	29-Mar	-
84	29-Mar	0.921	29-Mar	-
85	19-Apr	0.664	19-Apr	-
86	19-Apr	0.762	19-Apr	-
87	19-Apr	0.858	19-Apr	-
88	19-May	0.893	19-May	-
89	19-May	0.772	19-May	-
90	19-May	0.712	19-May	-
91	19-Jun	0.973	19-Jun	-
92	19-Jun	0.605	19-Jun	-
93	19-Jun	0.597	19-Jun	-
94	29-Jul	0.579	29-Jul	-
95	29-Jul	0.875	29-Jul	-
96	29-Jul	0.812	29-Jul	-
97	Aug-32	0.777	Aug-32	-
98	Aug-32	0.921	Aug-32	-
99	Aug-32	0.599	Aug-32	-
100	19-Oct	0.407	19-Apr	0.428
101	19-Oct	0.813	19-Oct	-
102	19-Oct	0.944	19-Oct	-
103	29-Dec	0.119	19-Oct	0.881
104	29-Dec	0.711	29-Dec	-
105	29-Dec	0.589	29-Dec	-
106	13-29	0.701	13-29	-
107	13-29	0.621	13-29	-
108	13-29	0.933	13-29	-
109	14-29	0.957	14-29	-
110	15-29	0.943	15-29	-
111	15-29	0.897	15-29	-

Table 4.3: Original Observations Rearrangement: Part 3

observation	actual type	prob	predicted type	prob
112	15-29	0.900	15-29	-
113	16-25	<0.001	19-Oct	0.824
114	16-25	0.013	19-Oct	0.455
115	18-29	1.000	18-29	-
116	18-29	1.000	18-29	-
117	18-29	1.000	18-29	-
118	21-25	0.387	21-25	-
119	21-25	0.114	25-32	0.635
120	21-25	0.896	21-25	-
121	25-32	0.960	25-32	-
122	25-32	0.826	25-32	-
123	25-32	0.985	25-32	-
124	26-29	0.995	26-29	-
125	26-29	0.955	26-29	-
126	26-29	0.965	26-29	-
127	27-25	0.955	27-25	-
128	27-25	0.823	27-25	-
129	27-25	0.892	27-25	-
130	28-25	0.959	28-25	-
131	28-25	0.936	28-25	-
132	28-25	0.871	28-25	-
133	29-25	0.802	29-25	-
134	29-25	0.981	29-25	-
135	29-25	0.827	29-25	-
136	30-25	0.938	30-25	-
137	30-25	0.947	30-25	-
138	30-25	0.852	30-25	-
139	31-51	0.842	31-51	-
140	31-51	0.814	31-51	-
141	31-51	0.839	31-51	-

Table 4.4: Original Observations Rearrangement: Part 4

regression methods.

Appendix A

R codes for model building.

```
data=read.csv(file.choose(),header=TRUE,sep=",")


library(nnet)
library(lmtest)
library(generalhoslem)


m1=multinom(ID~K,data)
m2=multinom(ID~Ti,data)
m3=multinom(ID~V,data)
m4=multinom(ID~Cr,data)
m5=multinom(ID~Mn,data)
m6=multinom(ID~Ni,data)
m7=multinom(ID~Cu,data)
m8=multinom(ID~Zn,data)
m9=multinom(ID~As,data)
m10=multinom(ID~Se,data)
```

m11=multinom(ID∼Br,data)

m12=multinom(ID∼Rb,data)

m13=multinom(ID∼Sr,data)

m14=multinom(ID∼Hg,data)

m15=multinom(ID∼Ga,data)

m16=multinom(ID∼Pb,data)

m17=multinom(ID∼Nd,data)

m18=multinom(ID∼Lu,data)

m19=multinom(ID∼Ta,data)

m20=multinom(ID∼Ln,data)

lrtest(m1)

lrtest(m2)

lrtest(m3)

lrtest(m4)

lrtest(m5)

lrtest(m6)

lrtest(m7)

lrtest(m8)

lrtest(m9)

lrtest(m10)

lrtest(m11)

lrtest(m12)

lrtest(m13)

lrtest(m14)

lrtest(m15)

lrtest(m16)

```
lrtest(m17)
lrtest(m18)
lrtest(m19)
lrtest(m20)

datadd=read.csv(file.choose(),header=TRUE,sep=",")

test=multinom(ID~K+Ti+V+Cr+Mn+Ni+Cu+Zn+
As+Br+Rb+Sr+Hg+Ga+Nd,data=datadd)
summary(test)
lrtest(test)

z=summary(test)$coefficients/summary(test)$standard.errors
z
p=(1 - pnorm(abs(z), 0, 1)) * 2
p

summary(test)$coefficients
summary(multinom(ID~K+Ti+V+Cr+Mn+Ni+Cu+Zn+
Br+Rb+Sr+Hg+Ga+Nd,data=datadd))$coefficients
summary(multinom(ID~K+Ti+V+Cr+Mn+Ni+Cu+Zn+
As+Rb+Sr+Hg+Ga+Nd,data=datadd))$coefficients
summary(multinom(ID~K+Ti+V+Cr+Mn+Ni+Cu+Zn+
As+Br+Sr+Hg+Ga+Nd,data=datadd))$coefficients
summary(multinom(ID~K+Ti+V+Cr+Mn+Ni+Cu+Zn+
As+Br+Rb+Sr+Ga+Nd,data=datadd))$coefficients
```

```
summary(multinom(ID∼K+Ti+V+Cr+Mn+Ni+Cu+Zn+
As+Br+Rb+Sr+Hg+Nd,data=datadd))$coefficients
summary(multinom(ID∼K+Ti+V+Cr+Mn+Ni+Cu+Zn+
As+Br+Rb+Sr+Hg+Ga,data=datadd))$coefficients

fitted(test)
logitgof(datadd$ID,fitted(test),g=30)

model=test

predict(model,newdata=data[,2:21])
```

Bibliography

[1] Vaishya, R., Chauhan, M., & Vaish, A. (2013). *Journal of Clinical Orthopaedics and Trauma*. Bone cement, 4(4), 157163. http://doi.org/10.1016/j.jcot.2013.11.005

[2] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. (2013). *Applied Logistic Regression*. Retrieved from https://ebookcentral-proquest-com.ezproxy.wpi.edu

[3] Fagerland, M. W. (2009). *Performance of Significance Tests, with Emphasis on Three Statistical Problems in Medical Research*. Series of Dissertations Submitted to the Faculty of Medicine, No. 853, University of Oslo.

[4] Fagerland, M. W., Hosmer, D. W., and Bofin, A. M. (2008). *Statistics in Medicine* Multinomial goodness-of-fit tests for logistic regression models. 27, 4238-4253.